Security Solutions for Privacy Preserving Improved Data Mining

Marian STOICA, Silvia TRIF, Adrian VISOIU
Academy of Economic Studies, Bucharest, Romania
marians@ase.ro, silviatrif@gmail.com, adrian.visoiu@predictie.ro

Approaches of data analysis in the context of Business Intelligence solutions are presented, when the data is scarce with respect to the needs of performing an analysis. Several scenarios are presented: usage of an initial dataset obtained from primary data as a reference for the quality of the results, enriching the dataset through decoration with derived attributes and enriching the dataset with external data. Each type of dataset decoration is used to improve the quality of the analysis' results. After being subject to improvement using the presented methods, the improved dataset contains a large number of attributes regarding a subject. As some attributes refer to sensitive information or imply sensitive information about the subject, therefore dataset storage needs to prevent unwanted analysis that could reveal such information. A method for dataset partitioning is presented with respect to the predictive capacity of a set of attributes over a sensitive attribute. The proposed partitioning includes also means to hide the link between the real subject and stored data.
Keywords: *Business Intelligence, Data Mining, Security, Privacy, Dataset Partitioning, Secret Sharing*

1 Introduction

Business Intelligence (BI) helps the decision making process. It relies on various data to offer reports, estimations and support [1]. At the base of BI, there are mechanisms for data processing. The results of data analysis are highly dependent on the hypothesis made for the analysis, the quality of the data, the algorithms used for processing. There are areas where for a proper analysis, available data is not sufficient or there is room for improvement. Such cases are discussed in the following. Presented methods improve the results of analysis. Some derived information is sensitive. We take into account that unauthorized access to the dataset could trigger the case when analysis may disclose sensitive information about the subject represented by a certain instance in the dataset. Security must be assessed and a solution is needed to cover this risk.

2 Problem Formulation of Security Issues Generated by Improved Prediction Using Enlarged Datasets

Various situations arise when BI tries to solve difficult problems that heavily impact the organization. In telecom industry, efforts are made to predict and to prevent existing customers to migrate to another operator from the market. In the electronic commerce, efforts are made to predict what clients need and to send them incentives, offers and bonuses. In stock markets and currency markets, predictions are made to estimate quotes and exchange rates.

Business intelligence has its power based on several aspects:

- the available data used to derive useful information for the business;
- the instruments used to process the available data;
- the instruments used to present the results to the decision maker.

The available data are mainly taken from the records made by the organization, from internal documents such as contracts, orders, invoices etc., activity history. The available data depends on the degree of how much electronic support is used over manual operations. Ideally, data is stored in files in databases and readily available when needed.

The instruments used to process data are used to:

- transform primary data according to various needs;
- apply algorithms in order to obtain results.

DOI: 10.12948/issn14531305/17.3.2013.14

The power of such instruments is given by the flexibility in accessing heterogeneous data and the quality and variety of algorithms used to obtain results.
The instruments used to present the results to the decision maker have to hide the complex processing from the previous type of instruments and show useful information in a handful way. Reports have to be easy to understand and conclusions must be highlighted in order to bring business value to the organization.
A common case is when a certain variable Y is studied. Variable Y is under the influence of various factors having associated the corresponding X_1, X_2, ..., X_n variables. For these, data series are built and various algorithms may be applied to obtain results. For example, this pattern applies well to regression or classification problems used to solve problems like churn prediction, suggestion of similar products, exchange rate forecasts.
Taking into account this problem formulation, at organization level, several cases arise with respect to the available data for solving the problem.
A particularity of some organizations is the fact that available data are restricted to internal sources. An example would be a mobile network operator having prepaid subscribers. The mobile operator does not have any knowledge regarding the subscriber identity, personal information other than the service usage the subscriber does, which is recorded. If the service provider wants to address the churn prediction problem then it is limited to the information that can be withdrawn from the call detail records.
In this case, from the primary data which is represented by the call detail records, through querying, data series may be obtained for the X_1, X_2... X_n variables or attributes. Estimations are made, and results are obtained regarding either the estimation of variable Y, in case of regression analysis, or the classification of instances, if classification is employed. Using a quality indicator for the results a quantitative value Q_1 shows how well the used algorithm performs on the respective dataset.

In the same context of data restricted to internal sources, an improvement over the quality of the result consists of improving the dataset by decoration. Transformation of the dataset containing the data series for attributes X_1, X_2, ..., X_n, obtained from primary data, is done through various methods:

- applying simple operations between attributes, based on the principles of building statistical indicators [2], obtaining a new subset of attributes U1, U2, ..., Um; this is a convenient way as the newly derived attributes have actual statistical meaning which makes them easy to interpret and use; it was shown that the quality of the result Q2 improved over the use of the initial dataset when using a decorated dataset, even if the source is the same;
- a more general approach is obtained by generating new attributes through gene expression programming; a new subset U1,U2,.. Um is added to the existing dataset; an algorithm is run and a new quantitative result Q2 is obtained; the result is better than the result obtained from the estimation of the initial dataset, without any decoration, as Q2>Q1. In [3] are presented information about the gene expression programming.

A solution to the problem of estimating a variable or classifying an instance when little data is available is to extend the existing dataset with data series corresponding to external information. The improved dataset X_1, X_2, ..,X_n, U_1, U_2, ..U_m is extended with external data that is transformed and a new subset V_1, V_2, .., V_p is added to the initial dataset. On this improved dataset, estimations are made and a new result is obtained, having a corresponding quality Q_3. Ideally, Q_3 is better than Q_2 or Q_1 if the external information brought into the dataset is relevant for the problem, e.g. if the external data series is associated to a factor that influence variable Y, in case of regression analysis or is relevant attribute for classification. Ideally, the following relation must be accomplished: $Q_3>Q_2>Q_1$. Figure 1 depicts the three stages of making use of available data.

DOI: 10.12948/issn14531305/17.3.2013.14

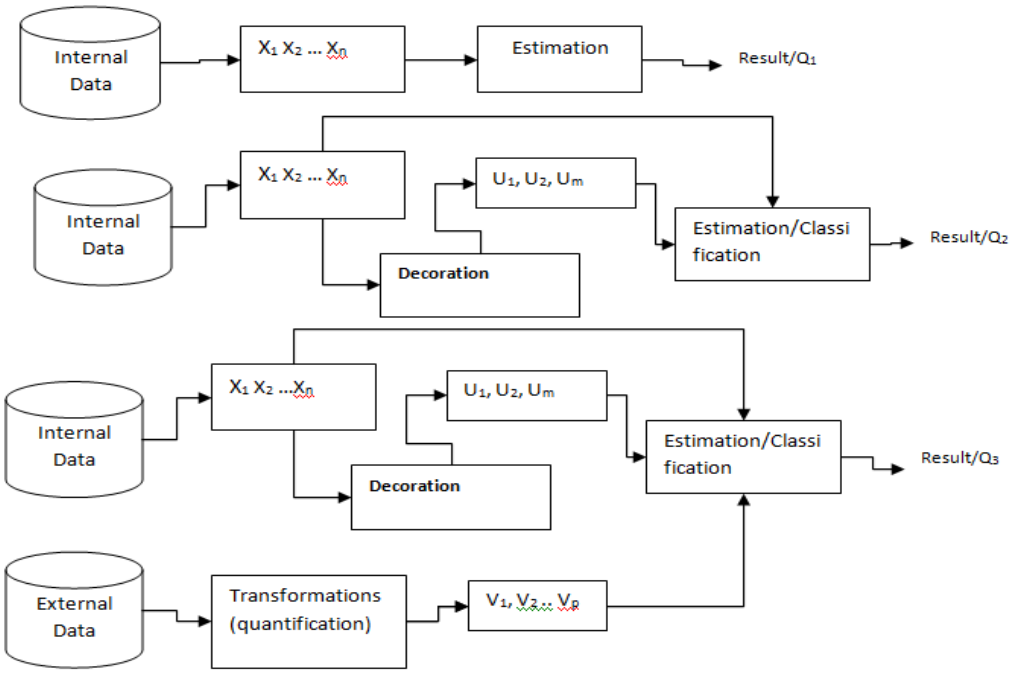

Fig. 1. Available data and means of result quality improvement

Each organization establishes the objectives. To reach the objectives, some decisions are made based on results given by the Business Intelligence solutions at organization level. In order to obtain the results, data has to be available to be processed by various algorithms. This way, the necessary data is assessed and design processes are executed to establish how to obtain the necessary data.

The result of employing presented dataset improvement has several consequences regarding the studied subjects:

- a relatively large number attributes is derived regarding the studied subject
- as the quality of the classification improves based on the existing attributes, extra attributes may be derived regarding the studied subject due to correlations between attributes; in [9] the author of the thesis demonstrates the possibility of deriving private attributes due to statistical relationship with public attributes. Therefore, if we take into account concrete cases, there exists the capacity of deriving private information about the customers just analyzing publicly available information by applying regression or classification algorithms.

Taking into account these peculiarities, a new problem arise regarding how derived data should be stored by the Business Intelligence storage engine in order to prevent an attacker to:

- link existing derived attributes with a certain person's identity
- estimate private information about a person based on the attribute information stored in the underlying data storage at the base of the Business Intelligence.

3 Solutions for Addressing Security Issues

The proposed solution to avoid estimating private attributes of a customer using existing records takes into account data partitioning. Approaches to dataset partitioning are found in [6] and [7]. For the problem of partitioning we consider the set of known attributes A_1, A_2, …, A_n. In order to protect a single private attribute A_j from being estimated by an attacker, the subset of known attributes that contribute to the correct estimation of A_j has to be determined. Also, attribute Aj may also indicate a public attribute but from another partition, such way gaining access to a partition, the attacker can't reconstruct the whole dataset. Considering that correlations or dependencies between attributes A_{i1}, A_{i2}, …, A_{ik} make estimation of Aj possible with a good fitness the aim is to segment these attributes into disjoint partitions in order for a

DOI: 10.12948/issn14531305/17.3.2013.14

certain partition to contain estimator attributes that are unable to estimate A_j with relevant precision. Therefore, if an attacker is able to gain access to a partition, it is unable to use attribute data found to estimate the protected attribute.

The steps for implementing this approach are:

1. depending on the application's objectives and available data, the solution analysts identify which sensitive attributes that are likely to be estimated based on the available data should be protected.
4. Also dependency between all attributes from the dataset has to be determined in case partitions have to be independent.
2. the list of dependency attributes A_{i1}, A_{i2}, …, A_{ik} that can be used to derive attribute A_j is obtained from the initial list A_1, A_2, …, A_n by applying an attribute selection algorithm that takes into account individual predictive ability, correlation or information gain in relation with the studied attribute A_j.
3. at dataset level, a dependency matrix is created as presented in
5. Table 1.

Table 1. Dependency matrix between attributes

	A_1	A_2	…	Aj	…	A_n	A_{p1}	A_{p2}	…	A_{pm}
A_1	$D_{1,1}$	$D_{1,2}$	…	$D_{1,j}$	…	$D_{1,n}$	$D_{1,p1}$	$D_{1,p2}$	…	$D_{1,pm}$
A_2	$D_{2,1}$	$D_{2,2}$	…	$D_{2,j}$	…	$D_{2,n}$	$D_{2,p1}$	$D_{2,p2}$	…	$D_{2,pm}$
…										
Ai	$D_{i,1}$	$D_{i,2}$	…	$D_{i,j}$	…	$D_{i,n}$	$D_{i,p1}$	$D_{i,p2}$		$D_{i,pm}$
…										
A_n	$D_{n,1}$	$D_{n,2}$	…	$D_{n,j}$	…	$D_{n,n}$	$D_{n,p1}$	$D_{n,p2}$		$D_{n,pm}$
A_{p1}	$D_{p1,1}$	$D_{p1,2}$	…	$D_{p1,j}$	…	$D_{p1,n}$	$D_{p1,p1}$	$D_{p1,p2}$		$D_{p1,pm}$
A_{p2}	$D_{p2,1}$	$D_{p2,2}$	…	$D_{p2,j}$	…	$D_{p2,n}$	$D_{p2,p1}$	$D_{p2,p2}$		$D_{p2,pm}$
Api	$D_{pi,1}$	$D_{pi,2}$	…	$D_{pi,j}$	…	$D_{pi,n}$	$D_{pi,p1}$	$D_{pi,p2}$	…	$D_{pi,pm}$
A_{pm}	$D_{pm,1}$	$D_{pm,2}$	…	$D_{pm,j}$	…	$D_{pm,n}$	$D_{pm,p1}$	$D_{pm,p2}$		$D_{pm,pm}$

In

Table 1 the dependency matrix between attributes is presented and the elements signify as follows:

A_1, …, A_n – available attributes in the dataset that is to be partitioned.

A_{p1}, …, A_{pm} – private attributes supposedly being estimated using A_1, …, A_n set. A_{p1}, …, A_{pm} are stored externally of the BI system, but are likely to be estimated using data from the BI system.

$D_{i,j}$ – dependency between attributes Ai and Aj from the same dataset; the value of Di,j is one if the attribute selection algorithm chose Aj as influential or correlated with Ai; A_j can be used to estimate A_i;

$D_{pi,j}$ – dependency showing A_j's influence or correlation with external and private attribute A_{pi};

In submatrix [$D_{p1,p1}$, $D_{pm,pm}$] dependencies are considered zero as dependency between external attributes is not in the scope of our system.

6. the following elements are taken into account the following:
   - consider NP the number of partitions to be created as an input, the partitions are the disjoint subsets P_i, with the property that P_1 U P_2 U … U P_{NP} = { A_1, …, A_n };
   - consider the attribute that needs to be protected from being estimated unwillingly, the secret attribute As, and the indicator $I_{QE}(A_s, P_i)$ representing the quality of the estimation of the secret attribute A_s using the attribute data from partition Pi.

DOI: 10.12948/issn14531305/17.3.2013.14

7. the best partitioning with respect to the protection of a single attribute As is P1, P2, …, PNP with the property that IQE(As, Pi) < ε, for all i=1,NP, where ε is an acceptance threshold and As does not belong to Pi;
8. in case more than one attribute is protected the set of secret attributes is As1, As2, …, Ask, then the partitioning P1, P2, …, PNP has to respect the property IQE(Asj, Pi) < ε, for all i=1,NP j=1,k and Asj does not belong to Pi.

In order to perform the actual partitioning, convenient number of partitions has to be chosen along with an appropriate indicator for the quality of the estimations.

For example, consider the simple case in which a single attribute is protected, the number of partitions is NP=2 and the indicator for the quality of estimations takes into account that the quality is proportional with the number of factors taken into account. In this case the partitioning will involve an heuristic algorithm, assuming influence over A_s is equally distributed among the factor attributes:

- determining the k attributes in the dependency list Ai1, Ai2, …, Aik;
- creating a partition P1 containing half of the attributes in the dependency list Ai1, Ai2, …, Aik/2 and half of the rest of the attributes in the initial attribute list;
- creating a partition P2 containing the second half of the attributes in the dependency list Ai(k/2)+1,…, Aik and the second half of the rest of the attributes in the initial attribute list.

The more general case involves many attributes, and many partitions available for storage. The number of possible combinations, the complexity of the evaluations makes difficult applying an exhaustive search method. Therefore we propose a genetic approach for the partitioning problem. The initial setup for the genetic approach is to map key concepts from the problem to be solved to key concepts from the evolutionary algorithm.

First of all, the solution that is searched is an allocation of attributes A_1, A_2, …, A_n to partitions P_1, P_2, …, P_{NP} such way the estimation capability for the protected attributes using a certain partition P_i is minimized. Therefore we consider an initial population of chromosomes that encode the allocation of attributes to partitions. In Table 2 the structure of a chromosome is presented.

Table 2. Structure of a chromosome used for genetic partitioning

G_1	G_2	…	G_i	…	G_{NP}
PA_1	PA_2	…	PA_i	…	PA_{NP}

The structure of the chromosome as presented in Table 2 contains a number of NP genes, G_1, G_2, …, G_{NP}, where G_i encodes the partition attribute A_i is allocated to, given by the value PA_i that takes one of the values {1, 2,.., NP}. An example of partitioning chromosome with five attributes and two partitions is presented in Table 3.

Table 3. Sample partitioning generated by chromosome encoding of length 5

G_1	G_2	G_3	G_4	G_5
2	1	1	2	1

The chromosome depicted in Table 3 encodes the following partitioning: for a set of five attributes, needed to be split into two partitions, the first partition contains attributes $P_1=\{A_2, A_3, A_5\}$, while the second partition is $P_2=\{A_1, A_4\}$.

In order to enable comparability of the solutions encoded by various chromosomes, each chromosome that encodes a partitioning solution is evaluated using an aggregated indicator built on top of the indicator used to assess the quality of estimation. For example, if the I_{QE} is additive, then a chromosome evaluation function E is:

$$E(\text{chromosome}) = \text{Sum } I_{QE}(A_s,P_i),\ i=1,NP$$

where

chromosome – the chromosome that is evaluated;

$I_{QE}(A_s,P_i)$ – the indicator showing the quality of the estimation of protected attribute A_s using attributes from P_i partition;

P_i – partition obtained selecting attributes corresponding to chromosome genes having

DOI: 10.12948/issn14531305/17.3.2013.14

PA_i=i.

Taking into account that we aim to limit estimation capacity for each partition, therefore the evaluation function is to be minimized.

An initial population of randomly generated chromosomes is evolved by applying genetic operators such as:

- selection – based on the value of the evaluation function, a proportion of the best chromosomes are retained to move to the next generation – i.e. chromosomes with low values of the evaluation function;
- cross-over – between randomly chosen chromosomes from the new generation genetic material is exchanged at random gene position; new chromosomes are obtained;
- mutation – randomly chosen chromosomes may have some values changed in order to introduce variability.

In Table 4 two chromosomes are presented in the state before cross-over. After cross-over at the position of gene G3, the resulting chromosomes are presented in Table 5.

Table 4. Partition encoding chromosomes before cross-over

Chromosome C_1				
G_1	G_2	G_3	G_4	G_5
2	2	2	1	2
Chromosome C_2				
G_1	G_2	G_3	G_4	G_5
1	1	2	2	1

Table 5. Partition encoding chromosomes after cross-over at position G3

Chromosome C_1'				
G_1	G_2	G_3	G_4	G_5
2	2	2	2	1
Chromosome C_2'				
G_1	G_2	G_3	G_4	G_5
1	1	2	1	2

As shown in Table 4, the initial chromosome C_1 encodes the partitions $P1_{C1}$ = { A_4} and $P2_{C1}$ = {A_1, A_2, A_3, A_5}. The initial chromosome C_2 encodes the partitions $P1_{C2}$ = {A_1, A_2, A_5} and $P2_{C2}$={A_3, A_4}. Following the cross-over at position G3, new individuals - chromosomes C_1' and C2' are obtained. Chromosome C1' encodes the partitions $P1_{C1'}$={ A_5 } and $P2_{C1'}$= {A_1, A_2, A_3, A_4}. Chromosome C_2' encodes the partitions $P1_{C2'}$ = {A_3, A_5 }. As seen the new individuals differ from their parents and therefore the different evaluations of these new chromosomes will bring improvement by creating better individuals.

Table 6 shows the effect of applying mutation operation on a chromosome.

Table 6. Mutation at position G1 in the chromosome

Chromosome C_1				
G_1	G_2	G_3	G_4	G_5
2	2	2	1	2
Chromosome C_1'				
G_1	G_2	G_3	G_4	G_5
1	2	2	1	2

By randomly changing genes inside chromosomes, variability is introduced leading to the creation of new individuals. Chromosome C_1 encodes the partitions $P1_{C1}$ = {A_4} and $P2_{C1}$ = {A_1, A_2, A_3, A_5}. After mutation, the new chromosome C_1' encodes the partitions P_1C_1' = {A_1,A_4} and $P2_{C1'}$ = {A_2, A_3, A_5}.

Starting with an initial population of chromosomes, randomly generated, genetic operations are applied in order to obtain new generations of chromosomes. After a threshold number of iterations the best fit individual is chosen with respect to the evaluation function. The partitioning encoded by the best chromosome is used to separate attributes into disjoint sets.

Once the partitioning problem is solved, an attacker that gains access to a certain partition is unable to estimate sufficiently well a protected attribute As in the absence of other attributes from other partitions. There remains only to add another layer of protection by anonymizing the data stored at partition level. This means that a certain record in the partition is identified by an identification attribute used to put in correspondence the record with the real entity that is characterized by the attributes stored at partition level. For example, in our case, the MSISDN identifies

DOI: 10.12948/issn14531305/17.3.2013.14

the row of data regarding stored attributes; also it belongs to a real person; the identification attribute needs to be used by the Business Intelligence system to retrieve and analyze data regarding a certain subscriber. If the attacker gains access to a partition and retrieves a row of data identified by a MSISDN, automatically the attacker gains access to information regarding the person that owns the MSISDN.

The proposed solution takes into account that there is a central node of Business Intelligence functionality where it is safe to use the identification attribute. Further than this point, at storage level layer, the identifier is considered as a secret and shared across partitions.

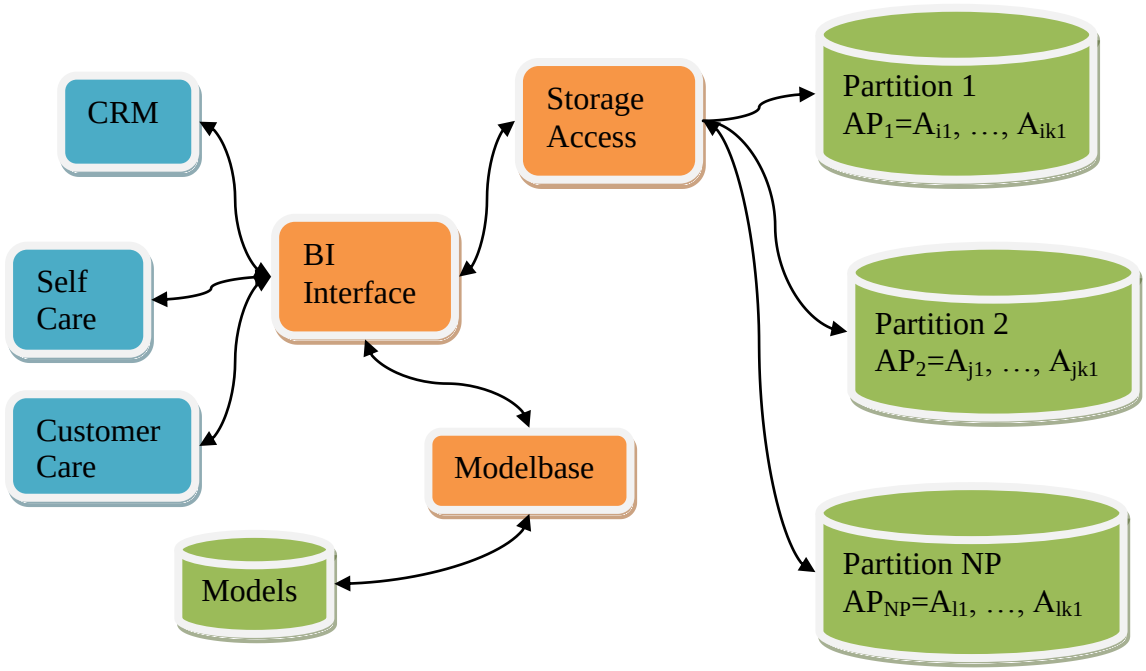

Fig. 2. BI interface integrated with client applications and using partitioned data

In the following, we will consider a common use case for the deployment of the Business Intelligence Solution. Taking into account the components presented in Fig. we discuss the case when the BI solution is deployed at a mobile operator type organization. The BI solution integrates with specific applications such as CRM or low level applications such as Customer Care or Self Care. These applications make use of the functionality provided by the BI layer. For example, an operator uses the Customer Care application to provision some bonuses for a certain user identified by a MSISDN. The application is developed such way to display to the operator information regarding the risk of becoming a churner. If the user is at risk then the operator will provision a bonus to stimulate the user to continue to use the mobile operator's services.

However, this action implies that the Customer Care application calls the Business Intelligence layer via a web service or other request/response protocol in order to obtain the risk measure. Also, the Customer Care application provides in the call the MSISDN of the user to retrieve information for. We consider that all the information about the user, information that is derived using the presented methods is part of a dataset stored separately. The matching between the requested information and the user is represented by the MSISDN. In order to prevent reverse matching and further association between the id of the data row and the user and also in order to prevent easy estimation of other attributes of the user based on the recorded ones, the dataset is partitioned into NP partitions. Since the MSISDN is used in clear inside the BI platform and inside the ecosystem of mobile

DOI: 10.12948/issn14531305/17.3.2013.14

operator's applications, on the storage layer, it is needed to have this value altered such way an attacker gaining access to a certain partition is unable to put in correspondence the id of a certain row of data with the MSISDN of the user of mobile operator's services. The solution for this problem is using a secret sharing scheme. We propose using Shamir's secret sharing scheme, as presented in [12]. We consider that the MSISDN is the secret to be hidden by sharing. An initial solution would be that in each partition, in each row, the MSISDN as identifier of the row is replaced by its part of the secret that corresponds to the partition, via Samir's transformation. However this does not cover all the security needs. If an attacker is interested in obtaining information regarding a certain MSISDN then he could easily use Samir's scheme to generate partitions of the secret which later are used to retrieve exact row from the partition. This issue is further addressed by the fact that the Storage Access layer does not need to know exactly the MSISDN but it can use any mapping or transformation given by the BI layer and known only by the BI layer which is considered trusted. The BI layer stores a secret key, K, which is used to encrypt the MSISDN. The encrypted message is then sent to the Storage layer which applies Shamir's scheme to obtain the identifiers for the user's data stored across the partitions. Storage Access layer becomes unaware of any relation between the real MSISDN and the received encrypted MSISDN, while at partition level each row id is even more displaced from the initial MSISDN and the real user identity it refers to.

In Fig. the sequence diagram shows the interactions between presented components in order to produce the desired result for the application's operator.

The operator accesses the page that request displaying of extra information regarding a certain subscriber identified by a MSISDN. Beside other information, the client application invokes a function on the Business Intelligence layer to obtain the churn risk associated to the user. In order to hide the sensitive value of MSISDN, the BI layer first encrypts the MSISDN using the secret key K, obtaining the encoded ENC_MSISDN. The BI layer has to retrieve the values for all attributes stored in the derived dataset in order to feed these values to a model and get a result.

DOI: 10.12948/issn14531305/17.3.2013.14

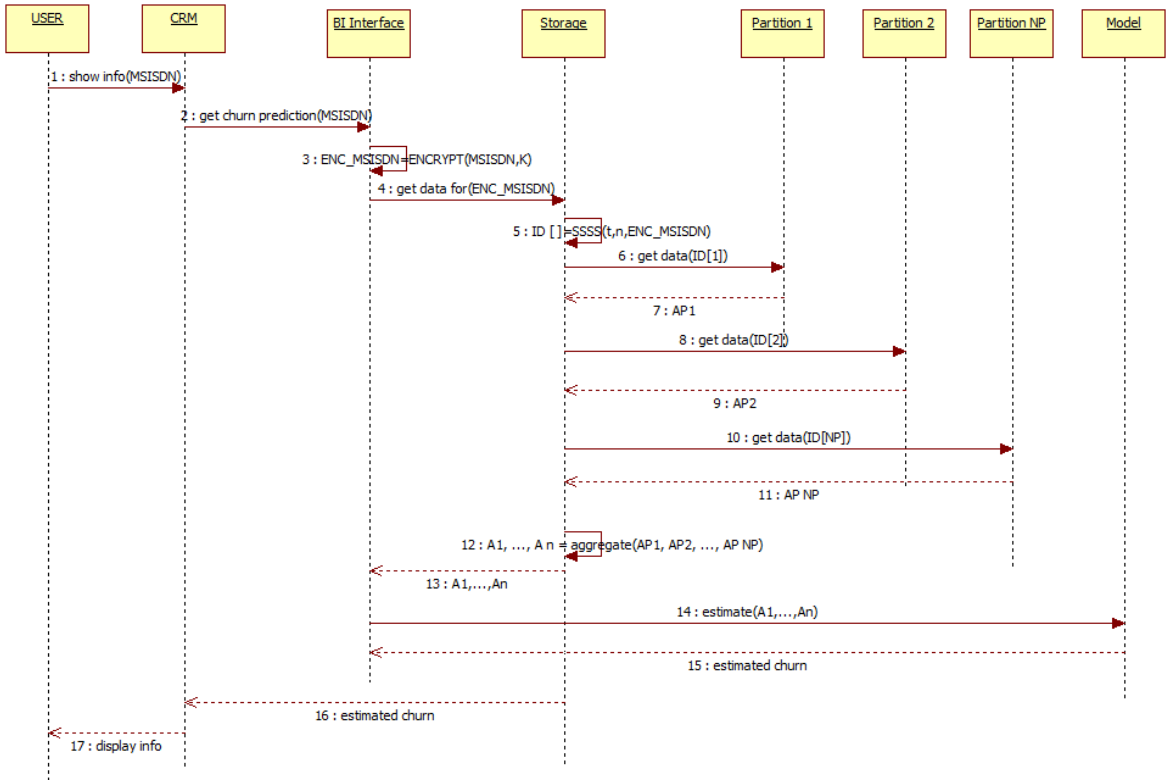

Fig. 3. Sequence diagram of the calls to securely retrieve user information

Therefore, BI layer calls Storage Access layer functionality to get data regarding the ENC_MSISDN identifier. The Storage Access layer has to query NP partitions of data. Each partition has the record regarding our user identified by a part of ENC_MSISDN. The Storage Access layer computes the list of identifiers, ID, using Shamir's scheme. For each partition that is managed by the Storage layer, a row with the data corresponding to a subset of attributes is retrieved based on the computed ID and data is returned to the Storage layer management component. Data returned from each partition is a row regarding a subset of attributes for the subject subscriber, and for partition *i* it is denoted AP_i. When data from all partitions have been retrieved, all the attributes are available regarding the user. Reunion of AP_1, AP_2, …, AP_{NP} is done by the Storage layer and itis exactly the list of values for the attributes A_1, A_2, …, An. Still at the level of Storage, no indication exists regarding the real user. The data is returned to the Business Intelligence layer. The BI layer then triggers modelbase functionality feeding the data to an estimation model in order to obtain an estimation of the churn risk. The obtained value is later passed in response to the client application to be included in the displayed report.

It is shown by the flow from Fig. and adjacent description that security is assured at different levels: by anonymizing partitioned data using secret sharing and also protecting attributes from being easily estimated by powerful data mining attacks.

3 Case Study

In the following we test our proposed solutions using a dataset. For comparison purposes the dataset is the same used in [10] and [11]. The dataset consists of a total of 18 attributes of which 4 are attributes from an initial dataset and the rest are obtained using the various improvement methods presented. The dataset includes 10694 rows and addresses churn prediction.

The considered attributes are regarding the usage of the service:

- usageX – amount of service consumed – number of calls – during week X
- diffwXwY – difference in usage between week X and week Y

DOI: 10.12948/issn14531305/17.3.2013.14

- rusgwX – relative usage in week X with respect to base week
- ardiffwXwY – absolute difference in relative usage between weeks X and Y
- pdiff – difference between week X and Y in percent
- state – detected state of the user – active or inactive indicating churning or not churning

As a step in the partitioning algorithm, dependency matrix for all attributes is computed as shown in Table 7 where attribute names are replaced with their index in the dataset.

Based on the nature of the attributes, it is decided which one is considered private. In the following we will consider that *state* attribute is to be protected. The hypothesis is made that there exists a model in the modelbase that takes all 17 factor attributes and is able to output an estimation of *state* attribute. *State* attribute is not intended to be stored in the dataset due to the fact that the model used for estimation had been fed with training data containing *state*, during training phase. All the subsequent estimations use the model until the next revision of the system including re-estimation of used models is done.

Using attribute selection algorithms, the dependency list for *state* attribute is obtained:

state <- {usagew1, usagew2, usagew3, usagew4, diffw4w3, rusgw1, rusgw2, rusgw3, rusgw4, ardiffw2w1, pdiffw3w1, pdiffw4w1}.

The remaining attributes are {diffw2w1, diffw3w2, ardiffw3w2, ardiffw4w3, pdiffw2w1}.

Using the whole training set, the quality of the estimation is given by the indicator that shows that 85.12% of the instances are correctly classified.

If partitioning is done into two partitions, each partition contains half of the dependency attributes and half of the rest of the attributes; since the attributes are related we will select them using a step of two.

Table 7. Dependency matrix between dataset attributes

Attribute	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1		1	1	1	1													
2	1		1	1	1	1			1			1	1					1
3	1	1		1		1	1	1		1								1
4	1	1	1				1	1			1	1		1				1
5	1					1						1						
6		1			1		1						1					
7			1			1								1				
8	1										1	1				1	1	
9		1										1	1					
10			1					1					1	1				1
11				1				1						1				
12					1			1	1						1			
13						1		1	1	1		1		1				
14							1			1	1		1					
15								1				1						
16								1		1								
17								1			1							
18	1	1	1	1			1	1	1	1	1	1				1	1	

The partitions are:

- P_1 = {usagew1, usagew3, diffw4w3, rusgw2, rusgw4, pdiffw3w1, diffw2w1, ardiffw3w2, pdiffw2w1};

DOI: 10.12948/issn14531305/17.3.2013.14

- P_2 = { usagew2, usagew4, rusgw1, ardiffw2w1, pdiffw4w1, diffw3w2, ardiffw4w3}.

Table 8. Partition evaluation

Partition	Evaluation (% of correctly estimated instances)
P1	66.93%
P2	58.56%
P1+P2 = full data	85.75%

The evaluation of the ability of attributes from individual partitions to estimate the protected attribute *state* is presented in Table 8.

As seen in Table 8 partitioning reaches its objective to prevent the ability to estimate a protected attribute if only a partition is available to an attacker. The quality of the estimations obtained only attributes from a certain partition is less than a threshold of 72% as indicator of usable estimations. In contrast, the whole dataset is able to give an accuracy of 86% which is considered excellent given the conditions of the analysis.

In order to ensure the anonymity of the data, the secret sharing scheme is applied at the provisioning of the data.

Table 9. Identifier splitting and reconstruction

Partitioning			
Partition P1		Partition P2	
ID	AP1	ID	AP2
…	…	…	…
46e32f3cc6e735	…	d49c0125d8983f	…
…	…	…	…
Reconstruction			
ID	A1, …, An = AP1+AP2		
7654321	…		

Given the Shamir secret sharing scheme for 2 partitions with a threshold of 2, given the identifier MSISDN=1234567, considering its encrypted form is ENC_MSISDN=7654321, then the sharing of ENC_MSISDN across the two partitions P1 and P2 is presented in Table 9 as well as the reconstruction of the identifier at Storage Access layer level.

As seen in Table 9 the records are impersonated. At partition level, the identifier of the row is only a share of the secret that is unable to reconstruct the secret. At Storage Access level, the secret is known but as the Storage Access does not know the key that generated the secret, the original identifier cannot be deducted.

5 Conclusions

The research in the series regarding decoration of datasets having little data available regarding the object of the analysis, with the purpose of improving data mining analysis has shown that an enlarged dataset is obtained that greatly improves the quality of the estimations done via regression or classification. Common applications used by large organizations need such results to help the decision making process and help the organization better understand its customers. However, it was shown that data mining methods are powerful enough to estimate or predict sensitive or private attributes based on the publicly available data. This constitutes a privacy issue that needs to be addressed. With regard to the proposed dataset improvement methods, we have shown that partitioning of datasets is a solution to prevent such attacks. The partitioning of data takes into account that a partition stores a subset of the attributes used to predict a sensitive attribute and an attacker gaining access to a certain partition is unable to estimate the sensitive attribute to a significant extent. Beside the data mining approach to the problem we addressed the anonymity of stored data, proposing a scheme to detach identification information from partitions from the actual entity that is subject to the analysis.

Since Business Intelligence solutions become more and more powerful and more data is available, such aspects presented in the paper must be taken into account accordingly.

DOI: 10.12948/issn14531305/17.3.2013.14

Acknowledgment
This work was co-financed from the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013, project number POSDRU/107/1.5/S/77213 „Ph.D. for a career in interdisciplinary economic research at the European standards".

References

[1] M. Chuah and K. Wong, "Business Intelligence - Solution for Business Development: Construct an Enterprise Business Intelligence Maturity Model (EBI2M) Using an Integration Approach: A Conceptual Framework", Business Intelligence Solution for Business Development, Intech, 2011, pp. 1- 14, ISBN: 978-953-51-0019-5.

[2] S. Trif and A. Vişoiu, "USSD based one-time password service", in Proc. of the 5th International Conference on Security for Information Technology and Communications 2012, Bucharest, pp 141 - 149.

[3] B. C. Ferreira, Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence 2nd Edition, Springer Publishing, May 2006.

[4] J. He, W. Chu and Z. Liu, "Inferring Privacy Information from Social Networks", in Proc. IEEE International Conference on Intelligence and Security Informatics, San Diego, California, 2006, pp. 154-165.

[5] M. Kantarcıoglu, J. Vaidya and C. Clifton, "Privacy Preserving Naive Bayes Classifier for Horizontally Partitioned Data", The International Journal on Very Large Data Bases, Vol. 17, No. 4, 2008, pp. 879-898

[6] J. Zhan, S. Matwin and Li Wu Chang, Data Mining: Foundations and Practice, chapter Privacy-Preserving Naive Bayesian Classification over Horizontally Partitioned Data, Springer Berlin Heidelberg, 2008, pp 528-538

[7] Z. Yang and R. Wright, "Privacy-Preserving Computation of Bayesian Networks on Vertically Partitioned Data", IEEE Transactions on Data Engineering, Vol. 18, No. 9, pp 1253-1264

[8] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations, Morgan Kaufmann, 2000

[9] S. Trif, A. Visoiu, "Using Data Mining Techniques to Infer Privacy Sensitive Attributes of Social Network Users", in. Proc. of the 6th International Conference on Security for Information Technology and Communications 2013, Bucharest, pp 313-320

[10] M. Stoica, S. Trif, A. Vişoiu, "Using External Datasources to Enrich Poor Datasets for Data Analysis", The 12th International Conference on Informatics in Economy Education, 2013, Bucharest, pag. 652-656

[11] S. Trif, A. Vişoiu, "Improving Churn Prediction in Telecom through Dataset Decoration", The 7th WSEAS International Conference on Computer Engineering and Applications CEA'13, 2013, Milano, pag. 223-228

[12] R. Baldoni, G. Chockler, "Collaborative Financial Infrastructure Protection: Tools, Abstractions and Middleware", Springer, 2012, ISBN 3642204198

Marian STOICA received his degree on Informatics in Economy from the Academy of Economic Studies, Bucharest in 1997 and his doctoral degree in economics in 2002. Since 1998 he is teaching in Academy of Economic Studies from Bucharest, at Informatics in Economy Department. His research activity, started in 1996 and includes many themes, focused on management information systems, computer programming and information society. The main domains of research activity are Information Society, E-Activities, E-Working, and Computer Science. The finality of research activity still today is represented by over 50 articles published, 9 books and over 20 scientific papers presented at

national and international conferences. Since 1998, he is member of the research teams in over 15 research contracts with Romanian National Education Ministry and project manager in 5 national research projects.

Silvia TRIF graduated the Faculty of Cybernetics, Statistics and Economic Informatics. She has a Master's Degree in Project Management. She is a PhD student of the Doctoral School of Bucharest Academy of Economic Studies in the field of Economic Informatics. Her interests are mobile applications, information security, web applications and project management. She has more the 15 articles in the fields of mobile applications security and Business Intelligence.

Adrian VISOIU graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2005. He holds a PhD diploma in Economics from 2009 and he worked at Bucharest Academy of Economic Studies as a teaching assistant until 2010. Currently he is an Associate Lecturer at the IT&C Security Master within the Department of Economic Informatics and Cybernetics at the Faculty of Cybernetics, Statistics and Economic Informatics from the Bucharest Academy of Economic Studies. He also works as a Senior Software Engineer for a software development company in the field of telecom. He is co-author of three books and over 40 articles in the field of model generation, data analysis, data mining with applications for software metrics, economic forecast, business intelligence and information security. His work focuses on data analysis, evolutionary algorithms and artificial intelligence.

DOI: 10.12948/issn14531305/17.3.2013.14